Initiation aux statistiques descriptives : cours

Mathieu Loiseau

Semestre 2, année universitaire 2009-2010

Table des matières

Ce document est en grande partie celui produit par Anahita Basirat, des actualisations ont cependant eu lieu çà et là.

Chapitre 1

Vocabulaire de base

1.1 Statistiques descriptives et inférentielles

Il s'agit d'organiser et résumer des observations. Le but est de décrire l'échantillon. Les statistiques inférentielles servent à étendre à la population les résultats ainsi obtenus. [Dancey & Reidy, 2007, p. 54]

1.2 Population

La population désigne un ensemble d'unités statistiques. Les unités statistiques, aussi appelées individus, sont les entités abstraites qui représentent des personnes, des animaux ou des objets. La statistique sert à décrire l'ensemble des unités statistiques qui composent la population.

1.3 Echantillon

Lorsque la population est trop importante, on étudie un échantillon, c'est-à-dire un sous-ensemble, beaucoup plus petit, de la population. L'échantillon doit être bien choisi pour pouvoir représenter la population.

1.4 Variables ou caractères statistiques

Un individu donné de la population peut être étudié selon certaines propriétés. Ces propriétes sont appelées caractères ou variables statistiques.

Exemple : Une étude sur les étudiants de l'université Stendhal peut porter sur les différentes variables : leur âge, leur sexe, leur nationalité, leur moyenne de l'année, etc.

1.4.1 Variables quantitatives

Une surface, un revenu moyen ou un âge sont des variables *quantitatives*. Elles peuvent être exprimées selon une unité de mesure et « peuvent être comparées entre elles, additionnées, faire l'objet de calculs de moyenne, [...] etc. » [Muller, 1973, p. 5].

a) Variables discrètes et continues

Sur un intervalle donné, les valeurs que peut prendre une variable quantitative *discrète* sont dénombrables (ex : nombre d'enfants d'un ménage). Au contraire, une variable quantitative *continue* peut prendre toutes les valeurs à l'intérieur d'un intervalle (ex : taille). En effet, entre une personne mesurant 160cm et 161cm, on peut imaginer une infinité de valeurs (ce qui n'existe pas entre 1 et 2 enfants par exemple). Ce sont la précision des instruments de mesure et les conventions qui font que la taille est traitée comme une variable discrète.

b) Classes

Pour pouvoir décrire des variables continues, il est parfois nécessaire de les « *discrétiser* », c'est à dire les répartir en *classes* : des intervalles de valeurs successifs. Les classes peuvent être définies en fonction du nombre de classe que l'on veut obtenir ou selon une amplitude fixe [1][Veysseyre, 2006, p. 9].

L'*amplitude* d'une classe est alors la différence entre la borne supérieure et la borne inférieure de l'intervalle défini : soit $[a; b[$ une classe d'une variable quantitative, on dit que $b - a$ est l'amplitude de cette classe.

Exemple : La taille d'un échantillon d'étudiants en CP peut être classée en moins d'un mètre, $[1; 1.20]$ mètre, plus de 1.20 mètre. L'amplitude de la seconde classe est 20 cm.

1.4.2 Variables qualitatives

Une variable *qualitative* est une variable qui ne prend pas de valeur numérique (elles ne répondent pas à une question « combien » mais à une question « est-ce que ») [Muller, 1973, p. 5].

Exemple : sexe, nationalité.

a) Modalités

Chaque variable qualitative a plusieurs *modalités*, ce sont l'ensemble des valeurs que la variable peut prendre.

Exemple : pour la variable sexe, les modalités sont masculin / féminin.

b) Variables ordinales

Une variable qualitative *ordinale* prend des valeurs qui sont ordonnées, hiérarchisées. On peut classer les modalités les unes par rapport aux autres mais on ne peut pas dire à partir de cet ordre de « combien » est la différence entre deux modalités.

Exemple : Les réponses à un sondage, du type « pas du tout », « un peu », « assez », « beaucoup » Véronis [2003].

1.4.3 Variables dépendantes et indépendantes

La statistique descriptive est un pré-requis à la statistique inférentielle, dans laquelle on applique des méthodes statistiques pour inférées des propriétés d'une population à partir d'un échantillon. La statistique permet de croiser des données pour tester des hypothèses. Dans ces cas là, une hypothèse consiste souvent à évaluer l'existence de l'effet d'une variable indépendante, que l'expérimentateur fera varier sur une variable dépendante qui sera évaluée. La description du comportement de ces deux variables sera primordial pour interpréter les résultats et leurs conséquences sur les hypothèses testées.

Exemple : On peut par exemple vouloir expliquer la taille des individus d'une population selon leur âge (exemple tiré de Wikipedia [2010]). Dans ce cas là, la variable indépendante est l'âge et la variable dépendante est la taille. Les deux variables pourront être décrites conjointement (cf. section 4.3 p. 16).

a) Variable parasite et variable de contrôle

Quand le but est de mesurer l'effet d'une variable indépendante sur une variable dépendante, il faudrait avoir des « groupes de sujets équivalents en tous points hormis les différences induites par les modalités de la variable indépendante. En d ?autres termes, il faudrait manipuler une variable indépendante et maîtriser toutes les autres. Les

1. Différentes méthodes ont été définies pour effectuer une « bonne discrétisation ». Voir par exemple : `http://www.info.univ-angers.fr/~gh/wstat/discr.php`.

variables indépendantes à maîtriser ou variable parasites (VP) sont très nombreuses et souvent inconnues. Ainsi on essaie de contrôler les variables parasites dont le chercheur sait ou présume l ?effet sur la variable dépendante. Les variables parasites fréquemment contrôlées sont :
– Les caractéristiques du sujets : le sexe, l ?âge, appartenance religieuse, politique ou culturelle ;
– Variable « expérimentateur » : lorsque plusieurs expérimentateurs recueillent des données, lorsque le sujet fait plusieurs tâches ou plus généralement appartient à plusieurs groupes expérimentaux.
Les variables parasites contrôlées s ?appellent *variables contrôles*. » Wikipedia [2010].

1.5 Effectif et fréquence

L'effectif d'une valeur donnée d'une variable est le nombre d'individus pour lesquelles la variable considérée prend la valeur en question. L'effectif total est la somme de tous les effectifs d'une variable.

La fréquence d'une valeur donnée est le rapport de l'effectif correspondant à l'effectif total. La fréquence totale est toujours égale à 1.

1.5.1 Exemple : une étude sur l'état matrimonial des salariés de la société X

– Population : salariés de la société X.
– Unité statistique (individu) : chaque salarié de la societé X.
– Variable (caractère) étudiée : état matrimonial avec 4 modalités : célibataire, pacsé ou marié, veuf, divorcé.
– Effectif : l'effectif de la modalité célibataire $= n_c$, pacsé ou marié $= n_{p/m}$, veuf $= n_v$, divorcé $= n_d$.
– Effectif total : N $= n_c + n_{p/m} + n_v + n_d$.
– Fréquence : fréquence de la modalité célibataire $= \frac{n_c}{N}$, pacsé ou marié $= \frac{n_{p/m}}{N}$, veuf $= \frac{n_v}{N}$, divorcé $= \frac{n_d}{N}$.
Fréquence totale $= \frac{n_c+n_{p/m}+n_v+n_d}{N} = \frac{N}{N} = 1$.

1.6 Effectifs cumulés croissants et décroissants

Note sur 20	< 5	$[5;10[$	$[10;12[$	$[12;15[$	$[15;17[$	$[17;20]$
Effectif	2	3	7	5	3	1
Fréquence	0.09	0.14	0.33	0.24	0.14	0.05
Effectif cumulé croissant	2	5	12	17	20	21
Effectif cumulé décroissant	21	19	16	9	4	1

TABLE 1.1: Exemple d'effectif cumulé : notes d'une population de 21 étudiants.

Quand les modalités ou les classes d'une variable sont rangées dans l'ordre croissant (resp. décroissant), les effectifs cumulés croissants (resp. décroissants) d'une valeur s'obtiennent en ajoutant à chaque effectif les effectifs des valeurs qui la précédent. Les fréquences cumulées s'obtiennent en divisant les effectifs cumulés par l'effectif total.

1.7 Série statistique

Une série statistique est la suite des observations d'une (ou plusieurs) variable(s), relevées sur les individus d'une population.

Exemple : Les notes des étudiants présentées dans le tableau 1.1 sous forme de classes. Elles auraient pu être représentées sous-forme d'une liste d'observations (notes) classées dans l'ordre croissant.

Note sur 20	2	4	5	7,5	9,5	10	10	10,5	11	11	11,5
	11,5	12	12	12,5	13	14	15	15,5	16,5	19	

TABLE 1.2: Exemple de série statistique classée dans l'ordre croissant.

Chapitre 2

Paramètres caractéristiques d'une variable : paramètres de position

Les paramètres de position (ou de tendance centrale) permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.

2.1 Mode

Pour une variable discrète, le mode est la modalité qui représente le plus grand effectif.

Exemple : sur la figure 2.1, le mode est « espagnol ».

LV2	allemand	allemand	allemand	anglais	anglais	**espagnol**	**espagnol**	**espagnol**
	espagnol	**espagnol**	italien	italien	italien	portugais	russe	tagalog

TABLE 2.1: Exemple de série statistique pour une variable discrète (qualitative ici).

Pour une variable quantitative continue, où la probabilité que chaque modalité n'apparaisse qu'une fois est supérieure (du fait qu'il existe une infinité de modalités), on a recours à des *classes modales* : c'est la classe dont l'effectif est maximum.

Exemple : Dans le tableau 1.1, la classe modale est la classe [10 ;12[.

2.2 Moyenne

2.2.1 Moyenne arithmétique

La moyenne arithmétique d'une série statistique est la somme des valeurs divisée par le nombre total des valeurs. Par exemple, la moyenne de l'année est la somme des notes de tous les examens divisée par le nombre d'examen. La moyen de X se calcule par $\overline{x} = \frac{x_1+x_2+\cdots+x_N}{N}$. Dans cette formule, $x_1, x_2, \cdots, x_N$ sont les notes et N est le nombre total des notes.

2.2.2 Moyenne pondérée

Lorsque les valeurs sont affectées de coefficients (ici d'effectifs), on parle de « moyenne pondérée » (voir tableau 2.2). La moyenne pondérée de X se calcule de la manière suivante :

$$\overline{x} = \frac{n_1x_1+n_2x_2+\cdots+n_Nx_N}{n_1+n_2+\cdots+n_N}$$

Dans cette formule, $n_1, n_2, \cdots, n_N$ sont les effectifs correspondants aux modalités $x_1, x_2, \cdots, x_N$.

Qualité de service	Effectif	Produit $n_i x_i$
1	1	1
2	3	6
3	5	15
4	2	8
5	1	5
total	12	35

TABLE 2.2: Moyenne de la variable qualité de service (Q_S) : $\overline{Q_S} = \frac{35}{12} = 2.9$

2.2.3 Propriétés

1. Considérons une série statistique S_1 de modalités $x_1, x_2, \cdots, x_N$ avec des effectifs $n_1, n_2, \cdots, n_N$ de moyenne $\overline{x}$ et la série statistique S_2 de modalités $y_1, y_2, \cdots, y_N$ avec des effectifs $n_1, n_2, \cdots, n_N$ telle que pour tout i appartenant à $\{1, 2, \cdots, N\}$: $y_i = ax_i + b$. Alors la moyenne de la série statistique S_2 est : $\overline{y} = a\overline{x} + b$.

   Exemple : La moyenne de notes d'une classe de 22 étudiants est 12.5. En ajoutant 0.5 point à toutes les notes, on obtient une moyenne de 13.

2. Soient S_1 et S_2 deux séries statistiques d'effectifs totaux respectifs N_1 et N_2 et de moyennes respectives $\overline{x_1}$ et $\overline{x_2}$. Alors la moyenne de la série S regroupant les deux séries S1 et S2 est : $\overline{x} = \frac{N_1\overline{x_1} + N_2\overline{x_2}}{N_1 + N_2}$. Cela revient à dire que la moyenne de plusieurs groupes correspond à la moyenne pondérée des moyennes pour chaque groupe (pondérée selon les effectifs des groupes).

   Exemple : La moyenne de notes d'une classe de 22 étudiants est 12.5 et celle d'une classe de 18 étudiants est 13.2. La note moyenne de ces deux classes est :
   $$\overline{x} = \frac{22 \times 12{,}5 + 18 \times 13.2}{22 + 18} = 12.81.$$

3. La moyenne n'est pas toujours un indicateur précis, elle est sensible aux valeurs extrêmes.

   Exemple : Dans un groupe de TD, 5 étudiants obtiennent $9,5$ au partiel et un étudiant $18,5$. La moyenne du groupe est de 11 [1] et pourtant seul un étudiant a validé le module correspondant...

2.3 Médiane

« La médiane (M_e) est la valeur, observée ou possible, dans la série des données classées par ordre croissant (ou décroissant) qui partage cette série en deux parties comprenant exactement le même nombre de données de part et d'autre de M_e » [Veysseyre, 2006, p. 15].

Exemple : Soit la série statistiques suivante : 15, 7, 22, 4, 12, 30, 9, 18, 6. Pour déterminer la médiane, il faut ordonner la série : 4, 6, 7, 9, 12, 15, 18, 22, 30. La médiane est le 12 car dans cette série, il y a 4 nombres inférieure et 4 supérieure de 12.

2.3.1 Calcul de médiane pour des données non réparties en classes

Si l'effectif total est impair (2n+1), la médiane est parfaitement déterminée : la modalité correspondant à n+1. Il s'agit d'une valeur observée.

Exemple : Dans le tableau 2.3, une étude sur le nombre d'enfant d'une échantillon de 51 individus $(2 \times 25 + 1)$ est présentée. La médiane est la modalité "1 enfant" qui correspond au foyer 26.

Si l'effectif total est pair (2n), on ne peut pas définir précisément la médiane : « on peut prendre pour valeur médiane, indifféremment l'une ou l'autre des valeurs centrales ou n'importe quelle valeur intermédiaire entre ces deux valeurs, par exemple, la moyenne arithmétique de ces deux valeurs, mais, dans ces conditions, ce n'est pas une valeur observée » [Veysseyre, 2006, p. 15].

1. $\overline{note} = \frac{5*9{,}5+18{,}5}{6}$

Nombre d'enfants	0	1	2	3	4
Effectif	20	16	10	5	0
Effectif cumulé croissant	20	36	46	51	51

TABLE 2.3: Calcul de médiane en utilisant les effectifs cumulés croissants : cas d'une variable discrète

Exemple : Une série représentant les notes d'une classe : 15, 7, 20, 4, 12, 20, 9, 18, 6, 4 (série ordonnée : 4, 4, 6, 7, 9, 12, 15, 18, 20, 20), l'intervalle médian est 9 et 12. Dans ce cas là, une acception de la médiane est $\frac{9+12}{2} = 10,5$. Il ne s'agit pas d'une valeur observée.

2.3.2 Calcul de médiane pour des données réparties en classe

Pour une variable continue, on détermine la classe médiane de même façon que pour une variable discrète en utilisant les effectifs cumulés. Exemple : dans le tableau 1.1, la *classe médiane* est la classe [10 ;12[. On détermine la médiane au sein d'une classe par l'interpolation linéaire.

Soit une étude sur la note d'une population de 50 étudiants (tableau 2.4) Levy [2010]. D'après la colonne "effectif cumulé", 18 personnes ont moins de 8 et 30 personnes ont moins de 12. La médiane se trouve donc dans l'intervalle [8 ;12[.

Notes	Effectifs	Effectifs cumulés
$[0; 5[$	10	10
$[5; 8[$	8	18
$[8; 12[$	12	30
$[12; 15[$	11	41
$[15; 20]$	9	50

TABLE 2.4: Calcul de médiane en utilisant les effectifs cumulés croissants : cas d'une variable continue

Sur la figure 2.1, les points A, X, B sont alignés et les droites AX, BX et AB ont le même coefficient directeur (la pente est la même). Le coefficient directeur d'une droite est déterminé par deux de ces points.

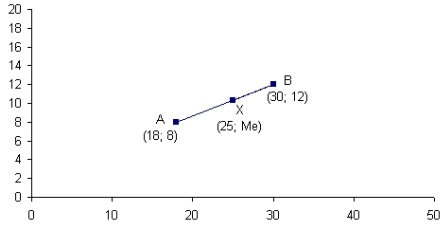

FIGURE 2.1: Calcul de médiane pour une variable continue. En abscisse : effectifs cumulés et en ordonnée : notes.

Le coefficient directeur de la droite AB se calcule par :

$$m = \frac{y_B - y_A}{x_B - x_A}$$

Pour trouver la valeur M_e, on peut calculer m_{AX} et m_{AB} et résoudre la régle de trois suivante :

$$m_{AX} = m_{AB} \text{ donc } \frac{M_e - 8}{25 - 18} = \frac{12 - 8}{30 - 18}$$

La médiane M_e est donc 10.33. Cela signifie que environ 50% des personnes ont eu moins de 10.33 et 50% plus de 10.33 (il s'agit à nouveau d'une valeur non observée, cependant dans ce cas précis nous n'avons pas le détail des valeurs observées puisque nous n'avons que les classes).

Chapitre 3

Paramètres caractéristiques d'une variable : paramètres de dispersion

Les paramètres de dispersion donnent des informations sur la répartition des valeurs autour de la valeur centrale de référence.

3.1 Etendue

L'*étendue* d'une série statistique quantitative est la différence entre la plus grande valeur de la variable (discrète ou continue) et la plus petite valeur. Exemple, dans le tableau 1.2, l'étendue est $19 - 2 = 17$.

3.2 Quantiles

Pour décrire des séries statistiques, le concept de médiane est adapté non plus pour séparer les mesures en 2 sous-ensembles, mais en k. On appelle ces mesures « *quantiles* ». Si $k = 4$ on parle de quartile.

3.2.1 Quartiles

Veysseyre définit les quartiles de la manière suivante :

> « Pour $k = 4$, les quantiles, appelés *quartiles*, sont trois nombres Q1, Q2, Q3 tels que :
> – 25% des valeurs prises par la série sont inférieures à Q1 ;
> – 25% des valeurs prises par la série sont supérieures à Q3 ;
> – Q2 est la médiane M_e ;
> – $Q3 - Q1$ est l'*intervalle interquartile*, il contient 50% des valeurs de la série.
> »
>
> [Veysseyre, 2006, p. 18]

Exemple : Levy [2010] : La *série ordonnée par ordre croissant* S a 12 termes :

$$S = \{11, 12, 13, 15, 16, 16, 17, 17, 18, 19, 20, 22\}$$

Rang	1	2	3	4	5	6	7	8	9	10	11	12
Série	11	12	13	15	16	16	17	17	18	19	20	22

TABLE 3.1: Calcule des quartiles

Un quart (25%) des données correspond à : $12 \times 0.25 = 3$. Le premier quartile est alors la plus petite valeur Q1 pour laquelle les valeurs de 3 termes de la série sont inférieurs ou égales à Q1. Le premier quartile est donc la valeur du 3ème terme de la série c'est-à-dire 13 [1].

1. Notez qu'il existe d'autres méthodes de calcules des quartiles et déciles : `http://www.math.unicaen.fr/irem/stat/quartile.pdf`

Trois quarts (75%) des données correspondent à : $12 \times 0.75 = 9$. Le troisième quartile est alors la plus petite valeur Q3 pour laquelle les valeurs de 9 termes de la série sont inférieurs ou égales à Q3. Le troisième quartile est donc la valeur du 9ème terme de la série c'est-à-dire 18.

L'intervalle interquartile est [13 ;18]. L'écart interquartile est 18-13 = 5.

3.2.2 Déciles

Quand $k = 10$, les quantiles sont appelés *déciles*.

- Le *premier décile* d'une série la plus petite valeur D1 des termes de la série pour laquelle au moins un dixième (10%) des données sont inférieures ou égales à D1 ;
- le *neuvième décile* (D9) d'une série est la plus petite valeur des termes de la série pour laquelle au moins neuf dixièmes (90%) des données sont inférieures ou égales à D9 ;
- l'*intervalle interdécile* est [D1 ;D9] ;
- l'*écart interdécile* est le nombre D9-D1.

Exemple : Levy [2010] : La *série ordonnée par ordre croissant* S a 11 termes :

$$S = \{1500, 1650, 1700, 1800, 1850, 2000, 2100, 2300, 2500, 2650, 2700\}$$

Rang	1	2	3	4	5	6	7	8	9	10	11
Série	1500	1650	1700	1800	1850	2000	2100	2300	2500	2650	2700

TABLE 3.2: Calcule des déciles

Un dixième (10%) des données correspond à : $11 \times 0.10 = 1.1$. Le premier décile est alors la plus petite valeur D1 pour laquelle les valeurs de 2 termes ($2 \geqslant 1.1$) de la série sont inférieurs ou égales à D1. Le premier décile est donc la valeur du 2ème terme de la série c'est-à-dire 1650.

Neuf dixièmes (90%) des données correspondent à : $11 \times 0.9 = 9.9$. Le neuvième décile est alors la plus petite valeur D9 pour laquelle les valeurs de 10 termes ($10 \geqslant 9.9$) de la série sont inférieurs ou égales à D9. Le neuvième décile est donc la valeur du 10ème terme c'est-à-dire 2650.

L'intervalle interdécile est [1650 ;2650]. L'écart interdécile est 2650 - 1650 = 1000.

3.3 Indicateurs de dispersion

3.3.1 Variance

La variance est un indicateur de la dispersion d'une série par rapport à sa moyenne. La définition de la variance d'une série statistiques est donnée par la formule :

$$V(x) = \frac{1}{N}\sum_{i=1}^{N} n_i(x_i - \overline{x})^2 = \frac{1}{N}\sum_{i=1}^{N} n_i x_i^2 - \overline{x}^2$$

$V(x)$ désigne la variance des n valeurs associées aux n unités statistiques de la population et $\overline{x}$ est la moyenne de ces unités statistiques.

3.3.2 Ecart-type

La définition de l'écart-type d'une série est donnée par la formule : $\sigma(x) = \sqrt{V(x)}$. Si l'écart-type est faible, cela signifie que les valeurs sont assez concentrées autour de la moyenne et si l'écart-type est élevé, cela veut dire au contraire que les valeurs sont plus dispersées autour de la moyenne.

Exemple : « Dans une usine, le fait d'avoir un écart-type aussi bas que possible peut constituer un objectif de contrôle de qualité. Soit une entreprise qui fabrique un certain composant et qu'un des éléments du contrôle de la qualité consiste à mesurer le diamètre du composant. Chaque composant aura donc son diamètre mesuré. On calculera ensuite le diamètre

moyen, puis l'écart-type. Si l'écart-type est faible, cela signifie que les pièces ont dans l'ensemble un diamètre proche de la moyenne, donc que leur diamètre se ressemble. À la limite, un écart-type nul signifie que toute les pièces ont le même diamètre. Inversement, plus l'écart-type est élevé, plus il y a de pièces dont le diamètre s'écarte de la moyenne et qui risque de ne pas cadrer avec le système auxquelles elles sont destinées. » [Mazerolle, 2010, p. 28–29]

a) Propriété

Considérons une série statistique S_1 de modalités $x_1, x_2, \cdots, x_N$ affectées des effectifs $n_1, n_2, \cdots, n_N$ d'écart-type $\sigma(x)$, et la série statistique S_2 de modalités $y_1, y_2, \cdots, y_N$ affectées des même effectifs $n_1, n_2, \cdots, n_N$ telle que pour tout i appartenant à {1, 2,..., N} : $yi = axi + b$. Alors : l'écart-type de la série statistique S_2 est : $\sigma(y) = |a|\,\sigma(x)$.

Chapitre 4

Représentations graphiques

4.1 Représentation d'effectifs/fréquences

Pour obtenir une meilleure compréhension de la répartition des valeurs d'une variable donnée d'une série, plusieurs types de diagrammes existent. Le choix de quel type de diagramme utiliser dépend directement du type de variable dont il s'agit.

4.1.1 Variables qualitatives

Diagramme en barre : dans ce diagramme, les modalités de la variable sont placées sur une droite horizontale et les effectifs (ou les fréquences) sont placés sur un axe vertical. La hauteur de la barre est proportionnelle à l'effectif (figure 4.1). Les barres ont une certaine épaisseur pour qu'il n'y ait pas de confusion avec les diagrammes en bâtons réservés à des variables quantitatives discrètes (figure 4.3).

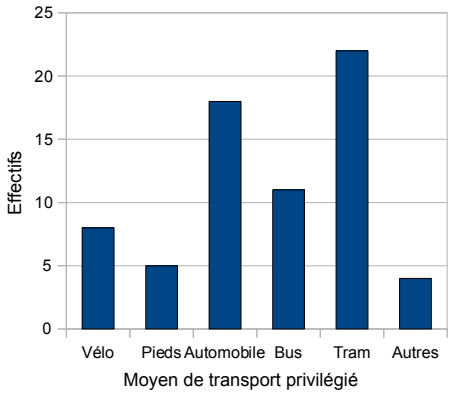

FIGURE 4.1: Diagramme en barre (Moyen de transport privilégié par une population d'étudiants – données fictives)

Diagramme circulaire ou camembert : L'effectif total est représenté par un disque. Chaque modalité est représentée par un secteur circulaire dont la surface (pratiquement : l'angle au centre) est proportionnelle à l'effectif correspondant (figure 4.2).

L'angle de chaque modalité se calcule par :

$$\frac{\text{effectif de la modalité}}{\text{effectif total}} \times 360°$$

4.1.2 Variables quantitatives

Diagramme en bâtons : Les valeurs discrètes x_i prises par les variables sont placées sur l'axe des abscisses, et les effectifs (ou les fréquences) sur l'axe des ordonnées. La hauteur du bâton est proportionnelle à l'effectif (figure 4.3).

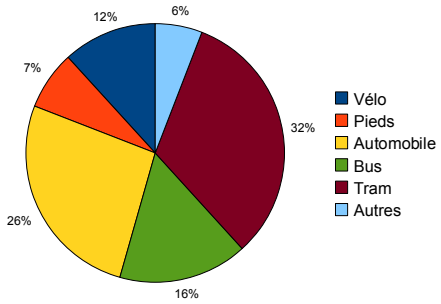

FIGURE 4.2: Diagramme circulaire (Moyen de transport privilégié par une population d'étudiants – données fictives)

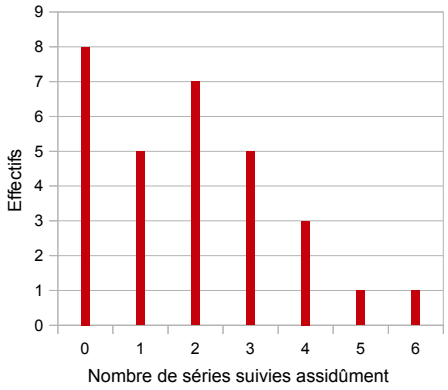

FIGURE 4.3: Diagramme en bâtons (Nombre de séries suivies assidûment par un population – données fictives).

Histogramme : on utilise l'histogramme pour les variables classées. C'est un ensemble de rectangles. Chaque rectangle est associé à une classe et il a une surface proportionnelle à l'effectif (ou fréquence) de cette classe.

- Amplitudes égaux : Si les classes ont la même amplitude, on reporte en ordonnée l'effectif (ou fréquence) des classes (voir figure 4.4 à gauche).
- Amplitudes diverses : si les amplitudes sont différentes, on reporte en ordonnée la *densité* d_i (effectif divisé par l'amplitude de la classe) pour que la surface de chaque rectangle soit proportionnelle à l'effectif (ou fréquence) (voir figure 4.4 à droite).

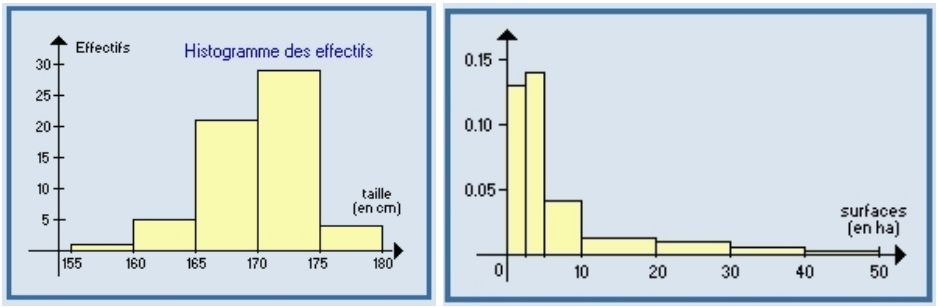

FIGURE 4.4: Exemples d'histogrammes. A gauche : des classes de même amplitude et à droite : des classes de différentes amplitudes (figures extraites de Janvier *et al.* [2002]).

4.2 Diagramme de Tukey

Les diagrammes de Tukey (ou boîtes à moustaches ou boîte à pattes) permettent de représenter sur une même figure des intervalles. En règle générale, on indique les valeurs extrêmes, les 1er et $(k-1)^e$ quantiles (Q1 et Q3 pour les quartiles, D1 et D9 pour les déciles, etc.) et éventuellement la médiane d'une série (figure 4.5).

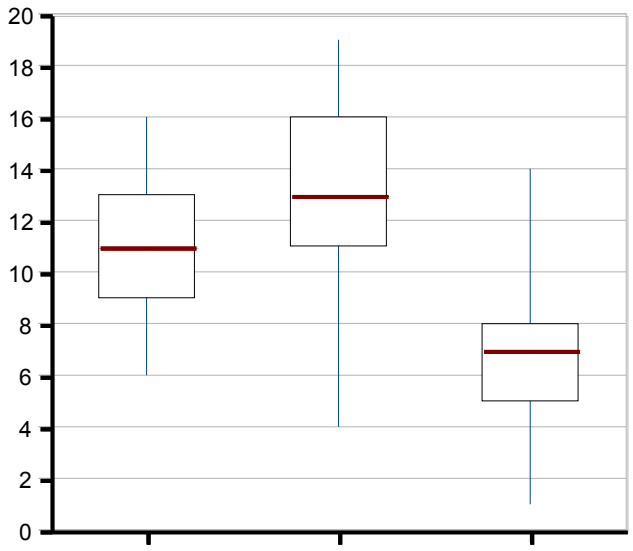

FIGURE 4.5: Exemple de boite à moustaches représentant les résultats d'une classe pour 3 devoirs successifs

Un tutoriel pour réaliser des diagrammes de Tukey dans OpenOffice est disponible à l'adresse suivante : `http://msp.aclyon3.free.fr/spip/IMG/pdf/bam_calc.pdf`

4.3 Statistique à deux variables

Lorsque l'on veux décrire deux variables, on pourra avoir recours à un nuage de points. La variable indépendante est représentée en abscisse et la variable dépendante en ordonnée. Reprenons l'exemple proposé dans la section 1.4.3 (p. 6). L'âge sera donc représenté en abscisse et la taille en ordonnée. La figure 4.6 représente un nuage de points pour ce problème.

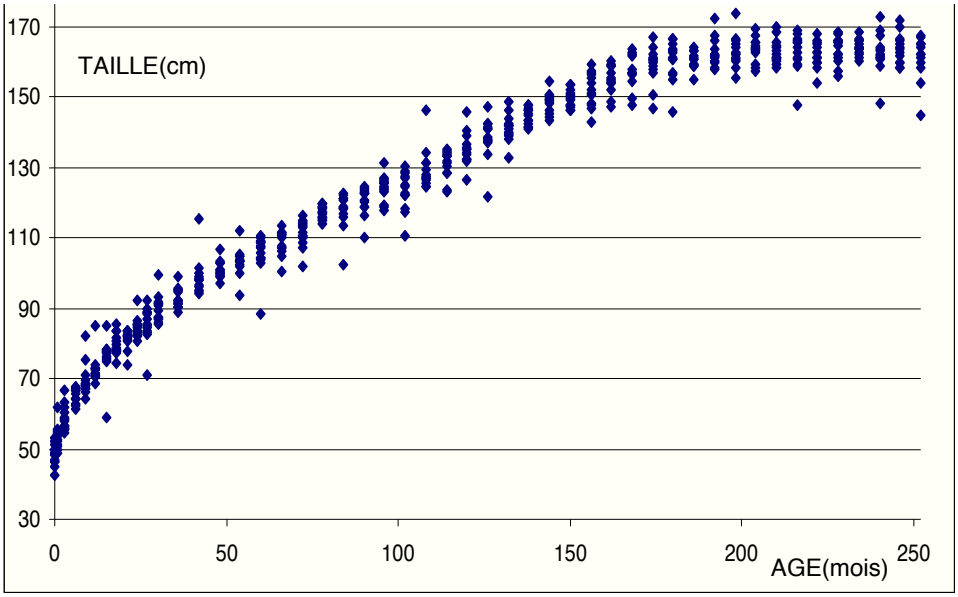

FIGURE 4.6: Taille en fonction de l'âge (filles), tiré de Gaudart [2010]

Chapitre 5

Quelques éléments de méthodologie d'enquête

Ces conseils sont issus, pour l'essentiel, de Ghiglione & Matalon [1978]; de Singly [1992, 2008] [1]. Il décrit la méthodologie de l'enquête du point de vue de la sociologie ; dans ce cadre elle « a pour objectif de produire de la connaissance, l'enquête par questionnaire ne se situe pas à un niveau exclusivement empirique. Elle engage un point de vue théorique, une vision du monde selon laquelle le social est déterminé socialement. » [de Singly, 1992, p. 22]. Il décompose une enquête en quatre temps qui s'influencent les uns et les autres :

> « Toute enquête a [...] des "biais" inévitables. Un questionnaire ne décrit jamais exhaustivement une pratique et, lorsqu'il approche trop précisément cette activité, les données seront ensuite regroupées, recodées pour éviter l'éparpillement et rendre possible l'analyse statistique. Le réel auquel renvoie l'objet de l'enquête est soumis à quatre principales transformations qui constituent les temps de la recherche : la délimitation opérée par la définition de l'objet ; la sélection des éléments jugés pertinents au travers des questions ; le tri par l'activité de codage et de recodage des informations recueillies ; la lecture d'une partie seulement des données. L'enquête est un long jeu de construction. »
>
> [de Singly, 1992, p. 21]

5.1 Avant le questionnaire

Selon [de Singly, 1992, p. 33] (ou [de Singly, 2008, p. 30]), l'avant-questionnaire doit permettre d'effectuer plusieurs tâches :
– lire ce qui a été écrit auparavant sur le thème étudié ;
– écouter ce que les acteurs sociaux disent de leurs propres pratiques ;
– élaborer progressivement une problématique.
C'est pourquoi, un questionnaire suit souvent une série d'entretiens, comme le font remarquer Ghiglione et Matalon : « Il est habituel de considérer qu'une enquête complète doit commencer par une phase qualitative, sous la forme d'entretiens non directifs ou structurés, suivie d'une phase quantitative, l'application d'un questionnaire à un échantillon permettant une inférence statistique au cours de laquelle on vérifie les hypothèses élaborées au cours de la première phase et on les complète par des renseignement chiffrés. » [Ghiglione & Matalon, 1978, p. 93]

Dans le cadre du projet que vous allez mener, il ne vous est pas demandé d'effectuer un travail de sociologue, à d'autant plus forte raison que vous n'aurez pas le temps de mener de front toutes ces activités. Il vous est cependant demandé de choisir une thématique pour vos questions et de l'expliciter dans votre rapport afin qu'elle soit claire pour le lecteur. Vous prendrez également soin de spécifier les objectifs de votre questionnaire entre (cf. [Ghiglione, 1987, p. 150]) :
– estimer ;
– décrire ;
– vérifier.

Nota Bene 5.1: Projet

1. Pour un point de vue moins étriqué que celui proposé ici, vous pouvez commencer par [Ghiglione & Matalon, 1978, pp. 93–138].

5.2 Types de question

5.2.1 Fait et opinion

Dans un questionnaire, il convient de faire la différence entre question de fait et question d'opinion. De Singly définit par l'exemple les deux notions [de Singly, 2008, p. 64] :
– « Combien de temps avez-vous passez à lire un livre hier ? » est une question de fait ;
– « Aimez-vous lire un livre ? » est une question d'opinion.
Il est important de savoir quel type de question vous êtes en train de poser. En effet, selon les objectifs de l'étude, certaines critiques peuvent être exprimées envers les questions d'opinion. Pour Bourdieu, « Un des effets les plus pernicieux de l'enquête d'opinion consiste précisément à mettre les gens en demeure de répondre à des questions qu'ils ne se sont pas posées » [Bourdieu, 2002, p. 226]. Nous pouvons également mentionner les commentaires de Lebaron :

> « Parmi les principales critiques que l'on peut adresser aux questions d'opinion, je citerai seulement ici :
> – le caractère très particulier des échantillons prélevés compte tenu de l'importance actuelle des *refus de réponse* ; les répondants effectifs sont de tous les répondants potentiels les mieux disposés à l'égard de l'enquête et du thème abordé ;
> – l'importance des *non-réponses* ; elle est réduite dans les conditions actuelles de saisie informatisée, mais elle fournit un indicateur de la distance à l'enquête qui caractérise beaucoup d'enquêtés ;
> – la signification fragile de nombreuses réponses qui sont en quelque sorte arrachées aux enquêtés, en particulier lorsque les questions ne font pas sens pour eux ou créent une situation artificielle (de type scolaire ou administrative)
> – d'une manière plus générale, toute forme de surinterprétation repose sur ce que l'on peut appeler l'*ethnocentrisme savant* : on prête aux enquêtés un raisonnement, des intérêts, une perspective qui sont ceux de l'enquêteur.
> »
>
> [Lebaron, 2006, p. 58]

Cependant, les questions de fait ne sont pas vierges d'approximations :
– « une question de fait ne doit pas demander aux personnes interrogées plus de précisions qu'elles ne peuvent en donner » [de Singly, 1992, pp. 63–64] ;
– « les individus sont beaucoup moins conscients de leurs pratiques qu'on ne le présuppose généralement. Aussi répondent-ils par approximation, même à des questions de faits » [de Singly, 1992, p. 64].

5.2.2 Questions ouvertes et fermées

Dans un questionnaire, on aura recours soit à des *questions ouvertes*, auxquelles l'utilisateur est libre de donner la réponse qu'il souhaite, soit à des *questions fermées*, pour lesquelles des réponses potentielles lui sont proposées.

Le choix du recours (principal) aux questions ouvertes ou fermées résulte en partie des objectifs de l'enquête menée.

> « Au-delà des problèmes de coût, la stratégie d'enquête face aux questions ouvertes et aux questions fermées renvoie donc à deux types d'enquête, l'une plus centrée sur le test d'hypothèses précises, l'autre plus attentive à la complexité du réel. Quelle que soit l'option retenue, le questionnaire comprend une majorité de questions fermées qui seront améliorées si ces deux procédures sont adoptées :
>
> 1. [entretiens préalables pour catégories mentales et mots utilisés] ;
> 2. Prévoir, dans la liste des réponses, une catégorie "autres" avec de la place pour ces réponses libres "imprévues".
>
> Lorsque, au moment du dépouillement, il y a beaucoup de réponses "autres", sont créées de nouvelles catégories. Mais cette technique n'est pas idéale, car rien ne garantit que la liberté de proposer des réponses "autres" puisse être appropriée également pour tous les individus. Cette solution est surtout intéressante lorsque les questionnaires sont remplis par les personnes interrogées elles-mêmes. »
>
> [de Singly, 1992, p. 69]

a) Terminologie : questions semi-fermées

Par convention, pour parler de questions fermées munies d'une catégorie « autres », nous parlerons de questions « semi-fermées ».

b) Difficultés liées aux questions ouvertes

Pour pouvoir effectuer une analyse quantitative des données, les questions ouvertes peuvent poser problème : « les personnes interrogées peuvent fournir des indications peu utiles. En effet, l'usage des questions ouvertes enseigne que nombre de réponses peuvent être floues, incodables. » [de Singly, 1992, p. 67]

En effet, une analyse statistique nécessite un codage des données, l'attribution de catégories selon les réponses apportées aux questions. Sans une préparation adéquate (cf. §5.1 p. 17), il n'est pas toujours possible de proposer des réponses aux enquêtés, sans introduire de lourds biais, cependant la décision d'avoir recours à des questions ouvertes doit être motivée.

5.2.3 Types de question et ordre

Selon le type de question, on aura intérêt à les placer en début ou en fin d'étude. Ghiglione et Matalon suggèrent par exemple de commencer par des questions ouvertes sous-peine de voir les sujets se contenter de réponses minimales quand ils parviennent à ces questions : « En principe, il y a avantage à regrouper toutes les questions portant explicitement sur le même thème. Sinon le sujet pourra avoir l'impression qu'on ressasse, qu'on lui demande quelque chose qu'il a déjà dit, qu'on tourne en rond. Toutefois, lorsqu'il s'agit de questions d'opinions, de préférences, d'attitudes, etc., il est parfois nécessaire de passer outre à cette recommandation pour éviter que les réponses ne soient biaisées par un souci de cohérence » [Ghiglione & Matalon, 1978, p. 99].

5.3 Recommandations

5.3.1 Favoriser l'expression personnelle

Pour plus de détails voir [de Singly, 2008, pp. 74–80].

Le choix des mots introductif est important, quand il s'agit d'une question d'opinion, il est pertinent de le faire paraitre dans la question par l'utilisation d'une brève proposition complémentaire du type « pensez-vous que » ou même « est-ce que », qui permettent à l'enquêté d'être moins soumis à une norme.

Pour que les enquêtés s'expriment librement, il est important de garantir leur « protection », par exemple en expliquant au début de la passation que les informations sont anonymes.

Il est conseillé d'assurer une certaine cohérence à l'enquête, en posant les questions qui s'y rapporte en premier, puis celles concernant les déterminants sociaux[2]. Enfin, s'il y a des questions « surprenantes », il est préférable de les conserver pour la fin.

5.3.2 Éviter les erreurs techniques

Éviter les erreurs techniques permettra d'améliorer la fiabilité des données obtenues[3].

- **Ne mettre qu'une seule question par question**, de plus, la question doit être précise, ne pas être sujette à interprétation ;
- ménager la mémoire des enquêtés : tenter, tant que faire ce peut, de limiter le nombre de réponses proposées à chaque question (moins important quand les individus interrogés gardent le questionnaire sous les yeux et peuvent relire les propositions) ;
- pour les questions d'opinion, proposer un continuum incluant réponses « extrêmes » (ex : « pas du tout d'accord, plutôt pas d'accord, plutôt d'accord, tout à fait d'accord »), il n'y a pas de consensus quant à savoir si une réponse « centriste » (centrale dans le continuum) doit être proposée, car une telle question peut servir de position de refuge ;
- éviter négations et doubles négations, qui rendent l'interprétation des questions (et ultérieurement des réponses...) difficile ;
- maitriser l'ordre des questions : l'ordre des questions peut influencer la manière dont les personnes y répondent[4].

2. Il s'agit de variables de contrôle (cf. § a) p. 6) fréquemment, pour ne pas dire systématiquement utilisées en sociologie, dont on sait qu'elles influencent souvent les réponses.

3. Pour plus de détails voir [de Singly, 2008, pp. 80–85].

4. Dans une étude réalisée en 1983 aux États-Unis, les réponses à la question « Pensez-vous que les États-Unis doivent accepter que des journalistes communistes originaires d'autres pays viennent ici et envoient chez eux des articles sur ce qu'ils ont vu ? » étaient radicalement différentes selon que la question « Pensez-vous qu'un pays communiste comme la Russie doit accepter des journalistes américains à venir là-bas et à envoyer aux États-Unis des papiers sur ce qu'ils ont vu ? » était posée avant ou après elle... [de Singly, 2008, p. 83]

5.4 La passation

Pour assurer la cohérence des résultats les conditions de passation doivent rester identiques :

> « [Cela] impose que ce soit le même questionnaire exactement qui soit posé à toutes les personnes interrogées. Une fois que le travail sur le terrain est commencé, il est donc exclu d'apporter des modifications quelconques à l'énoncé des questions ou à leur ordre, même su l'on est convaincu qu'il s'agit d'amélioration importantes, même si l'on s'est rendu compte d'erreurs graves. Dans ce dernier cas, il faudrait, dans l'idéal, recommencer entièrement l'enquête avec le questionnaire corrigé, et considérer les premiers sujets interrogés comme nuls. Si l'on ne dispose pas de suffisamment de moyens ou de temps pour cela, on poursuivra avec une version corrigée, mais en sachant que certaines questions n'avaient été posées qu'à une partie de la population, et que tous les dépouillements dans lesquels elles devaient intervenir ne pourraient se faire que sur ce demi-groupe de sujets.
>
> [...]
>
> C'est pour la même raison de maintien de la constance des conditions de passation qu'il ne faut pas que l'enquêteur se trouve obligé d'expliquer certaines questions à une partie des sujets. Le questionnaire doit être conçu de telle sorte qu'il n'y ait aucun besoin d'explications autres que celles qui sont explicitement prévues. »
>
> [Ghiglione & Matalon, 1978, pp. 95–96]

5.5 Analyse

5.5.1 Les non-réponses

L'analyse des non-réponses est crucial. Pour [Bourdieu, 2002, p. 225], le fait de les ignorer introduit un biais. Pour Mucchielli, les non réponses peuvent avoir différents sens ou différentes utilités [Mucchielli, 1990, p. 53] :

- non réponses dans des questionnaires remplis :
  - « Ignorance réelle du thème de la question par le sujet interrogé. Il se peut que ce soit justement un des objectifs de l'enquête que de mesurer si les individus du groupe représentant l'Univers de l'enquête,... savent ou non,... peuvent définir ou non,... comprennent ou non... quelque chose. »
  - « Refus de s'engager dans une réponse ferme ou dans les réponses prévues, et ceci peut avoir un sens. Dans certains cas, on peut y voir une attitude d'opposition. »
  - « Fuite de la réponse, car la question a éveillé inquiétude ou méfiance. Nous avons déjà signalé l'augmentation du nombre des "sans-réponses´´ aux questions d'opinion ou d'attitude, lorsqu'elles sont posées de façon directe. »
  - « Incompréhension de la question et refuge dans la non-réponse. »
- l'analyse du non renvoi du questionnaire :
  - sert à établir les caractéristiques de l'échantillon touché ;
  - peut mener à la détermination des caractéristiques des non répondants peut mener à la mise en place d'enquêtes complémentaires ;
  - peut servir de base à des calculs statistiques d'extrapolation.

5.5.2 Recodage des réponses

Une analyse qualitative et une catégorisation des réponses ouvertes est nécessaire pour effectuer un traitement quantitatif. Le recodage des réponses doit avoir lieu après avoir eu toutes les réponses. Dans le cas du recodage d'un type de réponse récurrent (par exemple un le codage d'une échelle allant de « pas du tout d'accord » à « tout à fait d'accord » en nombres de 1 à 4), il faut s'assurer que le même codage est utilisé pour toutes les questions qui y ont recours.

5.5.3 Présentation des tableaux

Comme pour les graphiques contenant plusieurs variables (cf. § 4.3 p. 16), la variable indépendante doit être disposée en ligne et la variable dépendante en colonne [5].

5. Pour plus d'informations, voir [de Singly, 2008, pp. 93–98].

Bibliographie

[Blanchet *et al.*, 1987] Alain Blanchet, Rodolphe Ghiglione, Jean Massonnat & Alain Trognon (1987). *Les techniques d'enquête en sciences sociales : observer, interviewer, questionner*. Dunod. ISBN : 2-04-016901-6.

[Bourdieu, 2002] Pierre Bourdieu (2002). L'opinion publique n'existe pas. Exposé à Noroit (Arras) (janvier 1972). Cf. *Questions de sociologie*, pages 222–235. Les éditions de minuit. ISBN : 2-7073-1825-6.

[Dancey & Reidy, 2007] Christine P. Dancey & John Reidy (2007). *Statistiques sans maths pour psychologues*. Ouvertures Psychologiques. De Boeck. ISBN : 978-2-8041-5384-7, ISSN : 1376-2273.

[de Singly, 1992] François de Singly (1992). *L'enquête et ses méthodes : le questionnaire*. sociologie 128. Nathan Université. ISBN : 209-190-567-4.

[de Singly, 2008] François de Singly (2008). *Le questionnaire*. L'enquête et ses méthodes. Armand Colin, 2e édition refondue. ISBN : 209-190-567-4.

[Gaudart, 2010] Jean Gaudart (2010). Principe de la régression linéaire.
url : `http://cybertim.timone.univ-mrs.fr/enseignement/doc-enseignement/statistiques/RL/docpeda_fichier`.

[Ghiglione, 1987] Rodolphe Ghiglione (1987). *Questionner*, pages 127–182. (Chapitre de Blanchet *et al.* [1987]).

[Ghiglione & Matalon, 1978] Rodolphe Ghiglione & Benjamin Matalon (1978). *Les enquêtes sociologiques : théories et pratiques*. Armand Colin. ISBN : 2-200-31046-3.

[Janvier *et al.*, 2002] Michel Janvier, Frédérique Kazi-Aoual, Mahmoud Hakim, Youssfi Elkettani, Martine Marco, Vincent Guijarro, Mireille Bachelot, Jean-Christophe Hamon, César Goemine, Thomas Meï, Henri Gwet & Aude Burgos (2002). *Statistique descriptive*. Techniques de la statistique.
url : `http://www.agro-montpellier.fr/cnam-lr/statnet/cours.htm`.

[Lebaron, 2006] Frédéric Lebaron (2006). *L'enquête quantitative en sciences sociales – recueil et analyse des données*. Psychologie Sociale. Dunod. ISBN : 2-10-048933-X.

[Levy, 2010] Jacques Levy. Math web. (2010).
url : `http://jellevy.yellis.net/`.

[Mazerolle, 2010] Fabrice Mazerolle (2010). *Statistique descriptive*, chapitre 3 : Statistiques permettant de résumer une série.
url : `http://www.mazerolle.fr/Statistique-descriptive/PlanHTML/statschapitre3.htm`.

[Mucchielli, 1990] Roger Mucchielli (1990). *Le questionnaire dans l'enquête psycho-sociale*. Formation permanente en sciences humaines. ESF, 9e édition. ISBN : 2-7101-0761-9.

[Muller, 1973] Charles Muller (1973). *Initiation aux méthodes de la statistique linguistique*. Numéro 32 *in* Unichamp. Champion, réédition 1992. ISBN : 2-85203-270-8.

[Véronis, 2003] Jean Véronis. Informatique et statistiques (chapitre 2.). (2003).
url : `http://sites.univ-provence.fr/veronis/cours/INFZ16/index.html`.

[Veysseyre, 2006] Renée Veysseyre (2006). *Statistique et probabilité pour l'ingénieur*. L'usine nouvelle. Dunod, 2e édition. ISBN : 2-10-049994-7.

[Wikipedia, 2010] Wikipédia (fr) (2010). Hypothèse statistique.
url : `http://fr.wikipedia.org/w/index.php?title=Hypoth%C3%A8se_statistique&oldid=45000342`.